

Publicación de datos del “Diccionario del Español Medieval electrónico” (DEMel)

Caroline Müller¹, Robert Stephan², Rafael Arnold³, Ulrike Henny-Krahmer⁴,
Karsten Labahn⁵ y Jutta Langenbacher-Liebgoth

Universidad de Rostock 2026

https://doi.org/10.18453/rosdok_id00005462

Licencia

CC BY-SA 4.0: Creative Commons Attribution-ShareAlike 4.0 International

<https://creativecommons.org/licenses/by-sa/4.0/>

Resumen

Bajo la dirección de Bodo Müller, se recopilaron materiales para un diccionario histórico del español medieval entre 1971 y 2007 en la Universidad de Heidelberg. Esta colección fue digitalizada e indexada en el marco del proyecto *Diccionario del Español Medieval electrónico* (DEMel), financiado por la Fundación Alemana de Investigación (DFG) entre 2016 y 2026. La base de datos resultante permite el acceso libre a más de 32 000 lemas con alrededor de 700 000 documentaciones procedentes de más de 600 obras, desde el siglo X hasta principios del siglo XV. En la presente publicación de datos, toda la información de la base de datos se publica en los formatos XML/TEI, RDF/TTL, JSON y SIARD, acompañada de una documentación introductoria, con el fin de facilitar su reutilización en diversos contextos de la investigación lingüística y las humanidades digitales.

Zusammenfassung

Unter der Leitung von Bodo Müller wurden Materialien für ein historisches Wörterbuch zum mittelalterlichen Spanisch zwischen 1971 und 2007 an der Universität Heidelberg zusammengetragen. Diese Sammlung wurde im Rahmen des von der Deutschen Forschungsgemeinschaft (DFG) zwischen 2016 und 2026 geförderten Projektes *Diccionario del Español Medieval electrónico* (DEMel) digitalisiert und erschlossen. Die daraus entstandene Datenbank ermöglicht den freien Zugang zu mehr als 32.000 Lemmata mit rund 700.000 Belegen aus über 600 Werken vom 10. bis zum frühen 15. Jahrhundert. In der vorliegenden Datenpublikation sind alle Informationen aus der Datenbank in den Formaten XML/TEI, RDF/TTL, JSON und SIARD veröffentlicht und mit einer einführenden Dokumentation versehen, um ihre Nachnutzung in verschiedenen Kontexten der linguistischen Forschung und der Digital Humanities zu erleichtern.

¹ <https://orcid.org/0000-0002-8591-7859>

² <https://orcid.org/0000-0001-7605-7415>

³ <https://orcid.org/0009-0003-1581-9086>

⁴ <https://orcid.org/0000-0003-2852-065X>

⁵ <https://orcid.org/0000-0002-8482-807X>

Abstract

Under the direction of Bodo Müller, materials for a historical dictionary of medieval Spanish were compiled at Heidelberg University between 1971 and 2007. This collection was digitized and indexed during the project *Diccionario del Español Medieval electrónico* (DEMel), funded by the German Research Foundation (DFG) between 2016 and 2026. The resulting database provides free access to more than 32,000 lemmas with around 700,000 attestations from over 600 works dating from the 10th to the early 15th century. In the present data publication, all information from the database is published in the formats XML/TEI, RDF/TTL, JSON, and SIARD, and is accompanied by introductory documentation to facilitate its reuse in various contexts of linguistic research and the digital humanities.

1. Descripción del proyecto y de los resultados

1.1. Descripción general

Este conjunto de datos forma parte del proyecto *Diccionario del Español Medieval electrónico* (DEMel), llevado a cabo en las Universidades de Rostock y Paderborn. El DEMel tiene su origen en el extenso archivo de datos recopilado en la Universidad de Heidelberg en el marco del *Diccionario del español medieval* (DEM), bajo la dirección de Bodo Müller. Iniciado en 1971, el DEM se convirtió en una obra lexicográfica de referencia para el estudio del español medieval, aunque quedó inconclusa cuando el proyecto fue suspendido a finales de 2007 por razones financieras. El archivo físico resultante comprende unas 865 000 fichas en papel, de las cuales alrededor de 700 000 documentan formas léxicas medievales con su contexto de uso, fuente y datación.

Tras la digitalización y el registro sistemático de los datos de las fichas, el DEMel ofrece acceso libre a más de 32 000 lemas⁶ con aproximadamente 700 000 documentaciones extraídas de más de 600 obras literarias y no literarias del siglo X hasta principios del XV. El proyecto fue financiado por la *Deutsche Forschungsgemeinschaft* (DFG, número de proyecto: 282533776)⁷ en tres fases entre 2016 y 2026.

1.2. Fases del proyecto

Fase 1 (2016–2020)

La primera fase se centró en la preservación y el registro digital del archivo, con el objetivo de asegurar a largo plazo los extensos datos de investigación del DEM y ponerlos a disposición de la comunidad científica. Para ello, las fichas fueron escaneadas e incorporadas a un repositorio, aproximadamente 700 000 fichas de documentaciones fueron indexadas manualmente y se desarrolló un portal web para la búsqueda.

Fase 2 (2020–2023)

La segunda fase tuvo como propósito profundizar en el procesamiento de los datos, ampliar las funcionalidades del portal web y abrirlo al público en general. Se incorporaron 43 000 imágenes adicionales con documentaciones secundarias, se precisaron las dataciones de las documentaciones y se añadieron aproximadamente 57 500 unidades léxicas pluriverbales a la base de datos. La bibliografía fue revisada en profundidad y enriquecida con datos de autoridad. El portal web recibió nuevas funcionalidades, entre ellas un editor y una función de retroalimentación.

⁶ El leuario también incluye aproximadamente 6 000 lemas de remisión.

⁷ <https://gepris.dfg.de/gepris/projekt/282533776>

Fase 3 (2023–2026)

La tercera fase completó la indexación del contenido. Los tres volúmenes impresos del DEM fueron retrodigitalizados, procesados mediante reconocimiento óptico de caracteres e integrados en el portal web. Se elaboró un corpus de textos del español medieval compuesto por 200 textos en formato XML/TEI con el fin de generar de ellos los contextos de las documentaciones. Finalmente, todos los datos fueron publicados en formatos estándar sostenibles e interoperables.

1.3. Equipo del proyecto

Dirección

El proyecto estuvo dirigido en su primera fase por Rafael Arnold (Universidad de Rostock, Departamento de Filología Románica), Jutta Langenbacher-Liebgott (Universidad de Paderborn, Departamento de Filología Románica), Robert Zepf (Biblioteca Universitaria de Rostock; hasta 2019) y Karsten Labahn (Biblioteca Universitaria de Rostock; desde 2019). En la segunda fase, la dirección correspondió a Rafael Arnold y Karsten Labahn y en la tercera fase a Rafael Arnold, Karsten Labahn y Ulrike Henny-Krahmer (Universidad de Rostock, Humanidades Digitales).

Coordinación

Ejercieron funciones de coordinación en distintos períodos del proyecto: Caroline Müller, Martin Reiter, Anna-Susan Franke, Stefan Serafin y Jan Reinhardt.

Desarrollo de la base de datos y las aplicaciones web

Robert Stephan, Caroline Müller y Karsten Labahn.

Colaboración estudiantil

Trinidad Almendra Arias Villalobos, Fiona Bochart, Livia Böhrs, Hanna Bott, Seda Coskuner, Hagar El Souhagy Louarmassi, Cenk Enyüce, Matthias Förster, Christian Gaponiuk, Elisabeth Gebauer, Anna Magdalena Goldmann-Klein, Linus Hauser, Mathilde Heeren, Eva Hinteresch, Dominika Jabłońska, Marie Kisser, Viacheslav Khodarev, Nele Sofie Naroska, Laura Noleppa, Regina Palamarcuk, Jördis Penning, Freyja Peters, Marcus Radloff, Erik Renz, Melanie Rulle, Laura Sackritz, Luise Scheibner, Sandra Schellnack, Daniel Scherma, Friederike-Sophie Schubert, Annely Seyfarth, Julia Siebert, Musab Torunoglu, Michelle Urbanski, Verena Welling y Katharina Wieben.

1.4. Resultados del proyecto

Portal web DEMel (<https://demel.uni-rostock.de>)

Aplicación web para la consulta de la base de datos, con funciones de búsqueda y filtrado. Los usuarios pueden descargar datos en distintos formatos o procesarlos de forma automatizada a través de interfaces de programación. El portal integra asimismo las imágenes de las fichas y los tres volúmenes impresos del DEM.

Software del portal DEMel

El código fuente está publicado en GitHub (https://github.com/ubrostock/demel_webportal) y en Zenodo (<https://doi.org/10.5281/zenodo.21873587>).

Presente publicación de datos

Contiene todos los datos del proyecto en distintos formatos, así como la documentación completa del proyecto:

Caroline Müller, Robert Stephan, Rafael Arnold, Ulrike Henny-Krahmer, Karsten Labahn y Jutta Langenbacher-Liebgott
Publicación de datos del “Diccionario del Español Medieval electrónico” (DEMel)
Universidad de Rostock 2026, https://doi.org/10.18453/rosdok_id00005462

Corpus de textos del español medieval

En la tercera fase del proyecto se elaboraron 200 textos en formato XML/TEI. Estos se encuentran disponibles como publicación de datos independiente en RosDok:

Caroline Müller, Robert Stephan, Rafael Arnold, Ulrike Henny-Krahmer y Karsten Labahn
Corpus de textos del “Diccionario del Español Medieval electrónico” (DEMel)
Universidad de Rostock 2026, https://doi.org/10.18453/rosdok_id00005463

Archivos de imagen de las fichas

Todos los archivos de imagen están archivados en el repositorio DORO de la Biblioteca Universitaria de Rostock.⁸

Archivos de imagen de los tres volúmenes impresos del DEM

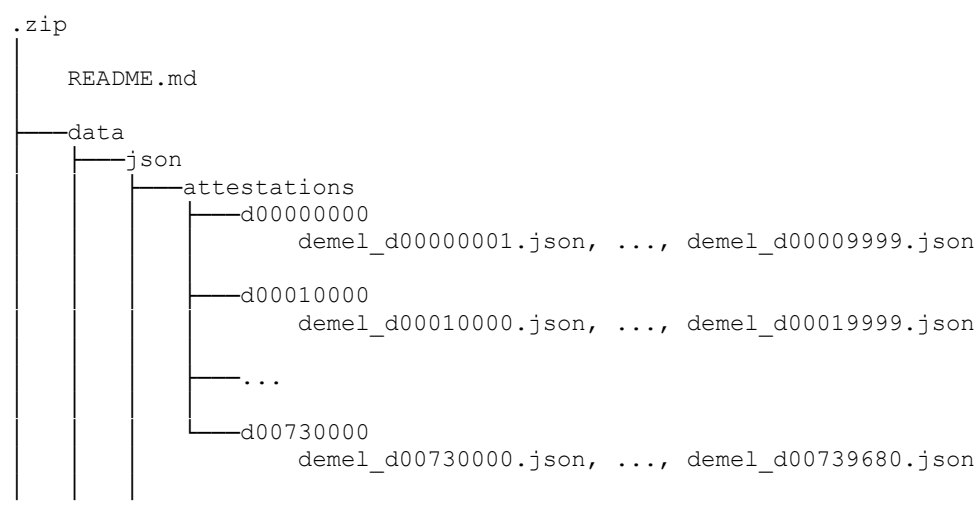
Los archivos de imagen son accesibles a través del repositorio RosDok de la Biblioteca Universitaria de Rostock: <https://purl.uni-rostock.de/rosdok/ppn1839132361>.⁹

1.5. Documentación adicional

Información más detallada sobre la historia del DEM, su archivo de datos, las tres fases del proyecto DEMel, los principios de registro y el uso del portal DEMel se encuentra disponible en español y en alemán en la carpeta `documentation/` de esta publicación.

2. Estructura de carpetas y archivos

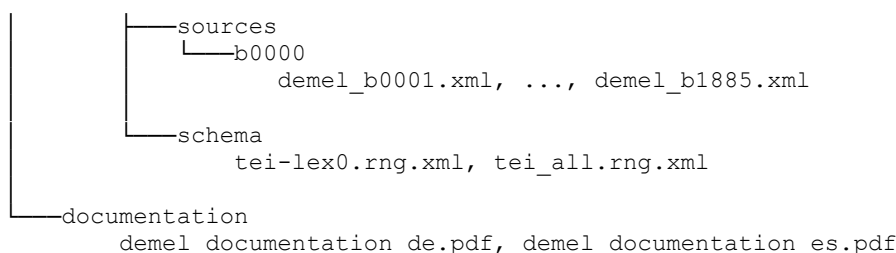
Este conjunto de datos ofrece los datos del proyecto en varios formatos estándar interoperables, organizados en dos carpetas principales: `data/` y `documentation/`.



⁸ Los enlaces permanentes a las gavetas que guardan las fichas figuran en los archivos JSON correspondientes de esta publicación de datos (`data/json/boxes/`). Las 482 gavetas originales con las fichas en papel se conservan en la Biblioteca Universitaria de Rostock.

⁹ Los enlaces permanentes a los tres volúmenes se incluyen en los archivos JSON correspondientes de esta publicación de datos (`data/json/boxes/`).

- boxes
 - k0000
 - demel_k0001.json, ..., demel_k0494.json
- lemmas
 - 100000000
 - demel_100000001.json, ..., demel_100009999.json
 - 100010000
 - demel_100010000.json, ..., demel_100019999.json
 - ...
 - 100040000
 - demel_100040000.json, ..., demel_100043177.json
- sources
 - b0000
 - demel_b0001.json, ..., demel_b1885.json
- rdf_ttl
 - demel_bibliography.ttl, demel_lexicon.ttl, demel_taxonomy.ttl
- attestations
 - d00000000
 - demel_d00000001.ttl, ..., demel_d00009999.ttl
 - d00010000
 - demel_d00010000.ttl, ..., demel_d00019999.ttl
 - ...
 - d00730000
 - demel_d00730000.ttl, ..., demel_d00739680.ttl
- lemmas
 - 100000000
 - demel_100000001.ttl, ..., demel_100009999.ttl
 - 100010000
 - demel_100010000.ttl, ..., demel_100019999.ttl
 - ...
 - 100040000
 - demel_100040000.ttl, ..., demel_100043177.ttl
- sources
 - b0000
 - demel_b0001.ttl, ..., demel_b1885.ttl
- relational
 - demel_db.model.svg, demel_db.schema.sql, demel_db.siard.zip
- tei_xml
 - demel_bibliography.tei.xml, demel_taxonomy.tei.xml
- lemmas
 - 100000000
 - demel_100000001.tei-lex0.xml, ..., demel_100009999.tei-lex0.xml
 - 100010000
 - demel_100010000.tei-lex0.xml, ..., demel_100019999.tei-lex0.xml
 - ...
 - 100040000
 - demel_100040000.tei-lex0.xml, ..., demel_100043177.tei-lex0.xml



2.1 Carpeta data/

Los datos se distribuyen en cuatro subcarpetas según el formato de serialización:

2.1.1 data/json/

Contiene los datos en un JSON específico del proyecto, pensado para su procesamiento automatizado y su integración en aplicaciones propias. Los datos están organizados en cuatro subcarpetas según el tipo de objeto:

- `attestations/` — Contiene todas las documentaciones registradas de las fichas. Los archivos se agrupan en subdirectorios de 10 000 entradas cada uno. Cada archivo corresponde a una documentación e incluye todos los datos de esta: la forma documentada, el lema y su categoría gramatical, la fuente y las imágenes de las fichas vinculadas. El conjunto abarca 704 088 registros. Los identificadores no están numerados de forma consecutiva.
- `boxes/` — Contiene los metadatos de las gavetas del fichero. Cada archivo corresponde a una gaveta y se encuentra en el subdirectorio `k0000/`. Los metadatos incluyen el PURL para acceder a cada gaveta con sus imágenes en el repositorio correspondiente.
- `lemmas/` — Contiene los objetos de lema (lemas y lemas de remisión). Los archivos se agrupan en subdirectorios de 10 000 entradas cada uno. Cada archivo corresponde a un lema. Se trata de 38 651 objetos de lema en total. Los identificadores no están numerados de forma consecutiva.
- `sources/` — Contiene todas las fuentes utilizadas para las fichas. Cada archivo corresponde a una fuente y se encuentra en el subdirectorio `b0000/`. El conjunto abarca 1 662 fuentes (fuentes primarias, secundarias, revistas y otras). Los identificadores no están numerados de forma consecutiva.

2.1.2 data/rdf_ttl/

Contiene los datos en formato RDF/Turtle, pensado para su uso en entornos de datos enlazados (Linked Data). Los archivos recogen únicamente los datos que pudieron modelarse mediante OntoLex Lemon¹⁰ o, en el caso de las fuentes, mediante schema.org¹¹, por lo que no necesariamente recogen la totalidad de los campos disponibles en otros formatos. Los archivos están organizados en las siguientes subcarpetas:

- `attestations/` — Contiene todas las documentaciones registradas de las fichas. Los archivos se agrupan en subdirectorios de 10 000 entradas cada uno. Cada archivo corresponde a una documentación. El conjunto abarca 704 088 registros. Los identificadores no están numerados de forma consecutiva.

¹⁰ <https://www.w3.org/2019/09/lexicog/>

¹¹ <https://schema.org/>

- `lemmas/` — Contiene los objetos de lema (lemas y lemas de remisión). Los archivos se agrupan en subdirectorios de 10 000 entradas cada uno. Cada archivo corresponde a un lema. Se trata de 38 651 objetos de lema en total. Los identificadores no están numerados de forma consecutiva.
- `sources/` — Contiene todas las fuentes utilizadas para las fichas. Cada archivo corresponde a una fuente y se encuentra en el subdirectorio `b0000/`. El conjunto abarca 1 662 fuentes. Los identificadores no están numerados de forma consecutiva.

Además, en el nivel raíz de esta subcarpeta se encuentran otros tres archivos:

- `demel_bibliography.ttl` — Contiene los metadatos de la bibliografía del proyecto con enlaces a todos los objetos de fuentes.
- `demel_lexicon.ttl` — Contiene los metadatos del proyecto con enlaces a todos los objetos de lemas publicados, a la bibliografía y la taxonomía .
- `demel_taxonomy.ttl` — Contiene la taxonomía del proyecto en formato SKOS.

2.1.3 `data/relational/`

Contiene la representación relacional de la base de datos, compuesta por tres archivos:

- `demel_db.siard.zip` — Base de datos completa en formato SIARD (Software Independent Archiving of Relational Databases)¹², estándar de preservación a largo plazo. Conforme al estándar, las tablas se convierten a formato XML y se guardan en un archivo en formato ZIP. El archivo fue generado con el programa “Siard-Suite” del Archivo Federal Suizo¹³.
- `demel_db.schema.sql` — Esquema de la base de datos en SQL.
- `demel_db.model.svg` — Diagrama visual del modelo de datos.

2.1.4 `data/tei_xml/`

Contiene los datos en formato XML conforme al estándar TEI (Text Encoding Initiative)¹⁴, pensado para su uso en entornos de humanidades digitales e investigación lexicográfica. En particular se usa el perfil TEI Lex-0¹⁵ para recursos lexicográficos. Los archivos se organizan en las siguientes subcarpetas y archivos:

- `lemmas/` — Contiene artículos básicos en formato TEI Lex-0. Cada archivo corresponde a un lema e incluye las documentaciones asociadas a este, por lo que no se publican las documentaciones como archivos individuales en formato TEI. Los archivos se agrupan en subdirectorios de 10 000 entradas cada uno. Los identificadores no están numerados de forma consecutiva.
- `sources/` — Contiene las fuentes en formato TEI. Cada archivo corresponde a una fuente y se encuentra en el subdirectorio `b0000/`. Los identificadores no están numerados de forma consecutiva.

¹² <https://dilcis.eu/content-types/siard>

¹³ <https://www.bar.admin.ch/bar/de/home/archivierung/tools---hilfsmittel/siard-suite.html>

¹⁴ <https://tei-c.org/>

¹⁵ <https://lex-0.org/>

- `schema/` — Esquemas de validación:
 - `tei-lex0.rng.xml` — Esquema RelaxNG para TEI Lex-0 del DARIAH Working Group on Lexical Resources.¹⁶
 - `tei_all.rng.xml` — Esquema RelaxNG general de TEI del TEI Consortium.¹⁷

Además, en el nivel raíz de esta subcarpeta se encuentran dos archivos:

- `demel_bibliography.tei.xml` — Contiene la bibliografía completa del proyecto ordenada por el tipo de fuente.
- `demel_taxonomy.tei.xml` — Contiene la taxonomía del proyecto en formato TEI.

2.2 Carpeta `documentation/`

Contiene la documentación del proyecto en PDF (`.pdf`), disponible en dos idiomas:

- Alemán: `demel_documentation_de.pdf`
- Español: `demel_documentation_es.pdf`

¹⁶ Esquema generado a partir del ODD correspondiente (<https://github.com/BCDH/tei-lex-0/blob/main/odd/lex-0.odd>)

¹⁷ Esquema generado a partir del ODD correspondiente (https://github.com/TEI/TEI/blob/dev/P5/Exemplars/tei_all.odd)